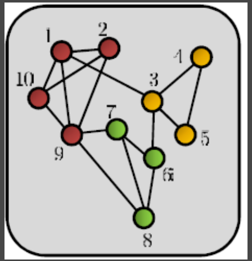

INFERENCE OVER NETWORKS

LECTURE #5: Logistic Regression

Professor Ali H. Sayed
UCLA Electrical Engineering

Course EE210B
Spring Quarter 2015

Proc. IEEE, vol. 102, no. 4, pp. 460-497, April 2014.
Foundations and Trends in Machine Learning, vol. 7, no. 4-5, pp. 311-801, July 2014.

Reference

Appendix G (Logistic Regression, pp. 777-780):

A. H. Sayed, ``Adaptation, learning, and optimization over networks," ***Foundations and Trends in Machine Learning***, vol. 7, issue 4-5, pp. 311-801, NOW Publishers, 2014.

Setting

Let $\boldsymbol{\gamma}_k$ denote a binary random variable whose value represents one of two possible classes, $+1$ or -1, depending on whether a feature vector $\boldsymbol{h}_k \in \mathbb{R}^M$ belongs to one class or the other. For example, the entries of $\boldsymbol{h}_k$ could represent measures of a person's weight and height, while the classes ± 1 could correspond to whether the feature $\boldsymbol{h}_k$ represents a male or a female individual. Logistic regression is a useful methodology for dealing with classification problems where one of the variables (the dependent variable) is binary and the second variable (the independent variable) is real-valued; this is in contrast to the more popular linear regression analysis where both variables are real-valued.

Logistic Function

When $\boldsymbol{\gamma}_k$ is a binary random variable, the relation between its realizations and the corresponding feature vectors $\{\boldsymbol{h}_k\}$ cannot be well represented by a linear regression model. A more suitable model is to represent the conditional probability of $\boldsymbol{\gamma}_k = 1$ given the feature vector $\boldsymbol{h}_k$ as a logistic function of the form [115, 234]:

$$P(\boldsymbol{\gamma}_k = +1 \mid \boldsymbol{h}_k) \;=\; \frac{1}{1+e^{-\boldsymbol{h}_k^{\mathsf{T}} w^o}} \tag{G.1}$$

Logistic Function

for some parameter vector $w^o \in \mathbb{R}^M$. Observe that regardless of the numerical values assumed by the entries of the feature vector $\boldsymbol{h}_k$, the logistic function always returns values between 0 and 1 (as befitting of a true probability measure) — see Figure G.1. Obviously, under the assumed binary model for $\boldsymbol{\gamma}_k$ and since the sum of the probabilities need to add up to one, it holds that

$$P(\boldsymbol{\gamma}_k = -1 \mid \boldsymbol{h}_k) \;=\; \frac{1}{1 + e^{\boldsymbol{h}_k^{\mathsf{T}} w^o}} \tag{G.2}$$

Logistic Function

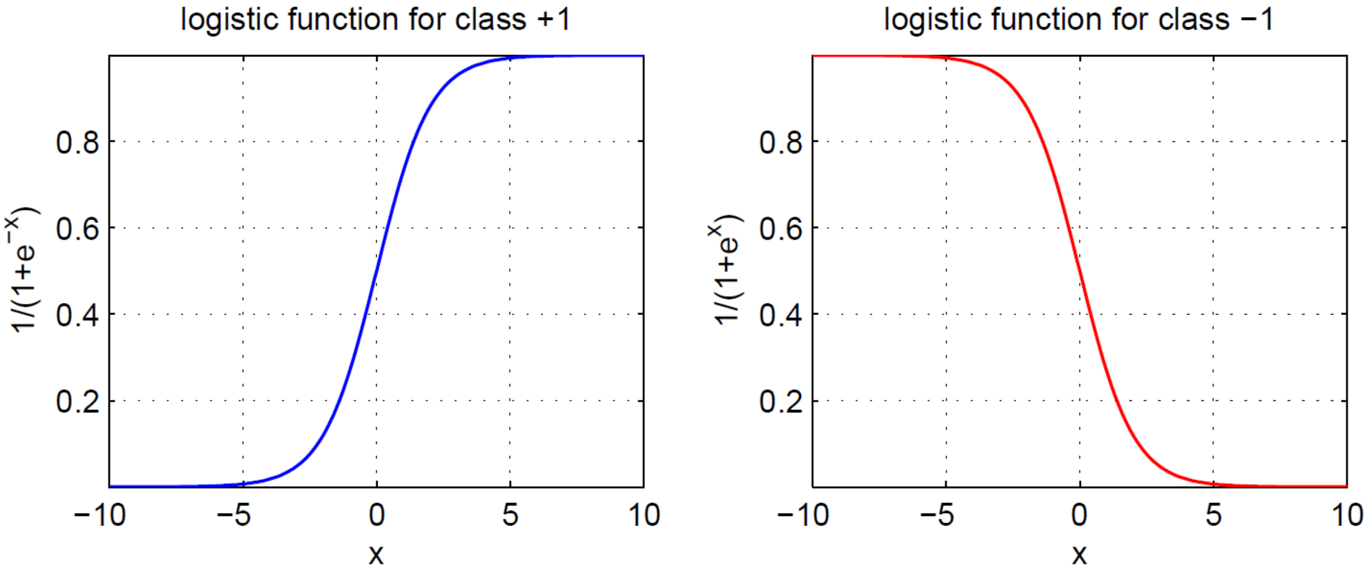

Figure G.1: Typical behavior of logistic functions for two classes. The figure shows plots of the functions $1/(1+e^{-x})$ (left) and $1/(1+e^{x})$ (right) assumed to correspond to classes $+1$ and -1, respectively.

Odds Function

We can group (G.1) and (G.2) into a single expression for the conditional probability density function (pdf) of $\boldsymbol{\gamma}_k$ and write:

$$p(\boldsymbol{\gamma}_k; w^o \mid \boldsymbol{h}_k) = \frac{1}{1 + e^{-\boldsymbol{\gamma}_k \boldsymbol{h}_k^{\mathsf{T}} w^o}} \tag{G.3}$$

with $\boldsymbol{\gamma}_k$ appearing in the exponent term on the right-hand side. This pdf is parameterized by w^o. In machine learning or pattern classification applications, one is usually served with a collection of training data $\{\boldsymbol{\gamma}_k, \boldsymbol{h}_k,\ k \geq 1\}$ and the objective is to use the data to estimate the parameter w^o.

Odds Function

Once w^o is recovered, its value can then be used to classify new feature vectors $\{\boldsymbol{h}_\ell\}$ into classes $+1$ or -1. This can be achieved, for example, by computing the odds of the new feature vector belonging to one class or the other. The odds function is defined as:

$$\text{odds} \triangleq \frac{P(\boldsymbol{\gamma}_\ell = +1 \mid \boldsymbol{h}_\ell)}{1 - P(\boldsymbol{\gamma}_\ell = +1 \mid \boldsymbol{h}_\ell)} \qquad \text{(G.4)}$$

Odds Function

For example, in a scenario where the likelihood that type +1 occurs is 0.8 while the likelihood for type −1 is 0.2, we find that the odds of type +1 occurring are 4–to–1, while the odds of type −1 occurring are 1–to–4. If we compute the log of the odds ratio, we end up with the so-called logit function (or logistic transformation function):

$$\text{logit} \triangleq \ln\left(\frac{P(\gamma_\ell = +1 \mid \boldsymbol{h}_\ell)}{1 - P(\gamma_\ell = +1 \mid \boldsymbol{h}_\ell)}\right) \tag{G.5}$$

Odds Function

There are at least two advantages for the logit representation of the odds function. First, in this representation of the odds, types $+1$ and -1 will always have opposite odds (i.e., one value is the negative of the other). And, more importantly, if we use the assumed model (G.1), then the logit function ends up depending linearly on w^o. Specifically,

$$\text{logit} \;=\; \boldsymbol{h}_\ell^\mathsf{T} w^o \tag{G.6}$$

In this way, we can assign feature vectors $\{\boldsymbol{h}_\ell\}$ with nonnegative logit values to one class and feature vectors with negative logit values to another class — see Figure G.2.

Pattern Classification

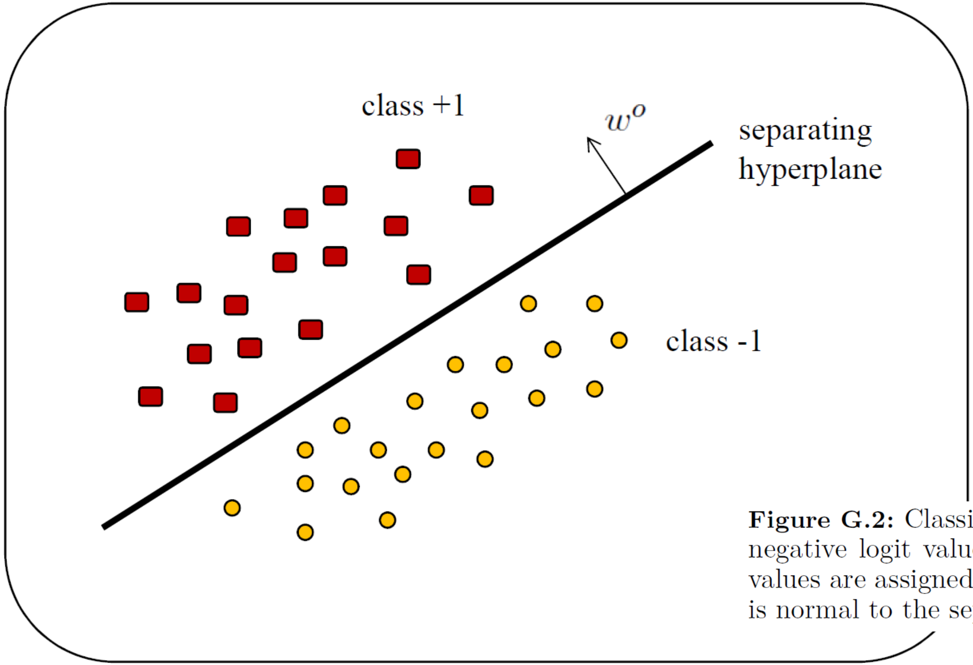

Figure G.2: Classification of feature vectors into two classes: data with non-negative logit values are assigned to one class and data with negative logit values are assigned to another class. The vector w^o defines the direction that is normal to the separating hyperplane.

Kullback-Leibler Divergence

To enable the above classification procedure, we still need to determine w^o. One way to estimate w^o is to fit into the training data $\{\boldsymbol{\gamma}_k, \boldsymbol{h}_k, k \geq 1\}$, a probability density function of the form:

$$p(\gamma_k; w \mid \boldsymbol{h}_k) \;=\; \frac{1}{1+e^{-\gamma_k \boldsymbol{h}_k^{\mathsf{T}} w}} \qquad \text{(G.7)}$$

for some unknown vector $w \in \mathbb{R}^M$ to be determined. This vector can be selected by minimizing the discrepancy between the above pdf and the actual pdf corresponding to w^o in (G.3).

Kullback-Leibler Divergence

A useful measure of discrepancy between two pdfs is the Kullback-Leibler (KL) divergence measure defined as [81]:

$$D_{\mathrm{KL}} \triangleq \mathbb{E}\left\{ \ln\left(\frac{p(\boldsymbol{\gamma}_k; w^o \mid \boldsymbol{h}_k)}{p(\boldsymbol{\gamma}_k; w \mid \boldsymbol{h}_k)} \right) \right\} \tag{G.8}$$

where the expectation is over the distribution of the true pdf. The expression on the right-hand side involves the ratio of two pdfs: one using the true vector w^o and the other using the parameter w. Minimizing over w leads to the optimization problem

Kullback-Leibler Divergence

$$\min_{w} \; -\mathbb{E} \ln p(\boldsymbol{\gamma}_k; w \mid \boldsymbol{h}_k) \tag{G.9}$$

or, equivalently,

$$\min_{w} \; \mathbb{E} \left\{ \ln \left[1 + e^{-\boldsymbol{\gamma}_k \boldsymbol{h}_k^{\mathsf{T}} w} \right] \right\} \tag{G.10}$$

which has the same form as the logistic regression cost function considered in the text — see, e.g., (2.9).

End of Lecture